GCN2CDD: A Commercial District Discovery Framework via Embedding Space Clustering on Graph Convolution Networks

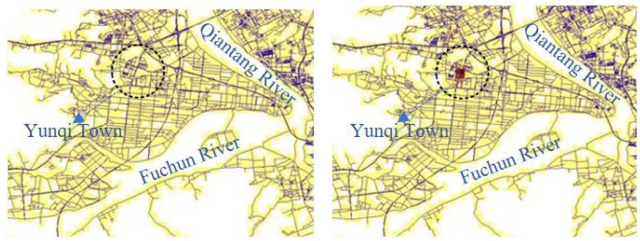
Guojiang Shen, Zhenzhen Zhao, and Xiangjie Kong, *Senior Member, IEEE*

***Abstract*—Modern enterprises attach much attention to the selection of commercial locations. With the rapid development of urban data and machine learning, we can discover the patterns of human mobility with these data and technology to guide commercial district discovery. In this article, we propose an unsupervised commercial district discovery framework via embedding space clustering on graph convolution networks to solve the problem of commercial district discovery. Specifically, the proposed framework aggregates human mobility features according to geographic similarity by graph convolution networks. Based on the graph convolution network embedding space, we apply hierarchical clustering to mine the latent functional regions hidden in different human patterns. Then, with the kernel density estimation, we can obtain the semantic labels for the clustering results to discover commercial districts. Finally, we analyze the multisource data of the Xiaoshan District and Chengdu City, and experiments verify the effectiveness of our framework.**

***Index Terms*—Commercial district discovery, embedding space, graph convolution networks (GCNs), human mobility.**

I. Introduction

ONE commercial district plays an important role in urban cultural and economic development. It directly determines the level of the city's development. Metropolitan have above average consumption power, and they expect that the city can provide satisfactory services. However, where to build a commercial site in a city is a troublesome problem. Commercial districts mainly own high popularity and convenient transportation. Traditionally, merchants will do surveys on the surrounding facilities of one candidate district and rely on their experiences to tell whether it can become a mature commercial center. The task is very time-consuming and labor-intensive. Luckily, with the development of industrial technology, many data sensors are placed in one city and all taxies are almost disposed of a GPS device, which can reflect the location of the taxi. Furthermore, the mature geographical information system will tell us about the distribution of points of interest (POIs). Thus, urban heterogeneous data describe how a city operates, which will open up a new opportunity to tackle the problem of commercial district discovery. Moreover, urban dwellers are active in different regions at different times in the city, and there is a latent consensus among them. Therefore, we can analyze different human mobility patterns to guide the discovery of commerce. Taking an intuitive example to elaborate on urban human mobility, there is an international conference on October 11, 2017, in Yunqi, a beautiful town in Hangzhou, China. We can see the change in taxi trajectory data's density obviously, as shown in Fig. 1.



(a) (b)

Fig. 1. Heat map of the taxi trajectory data in Yunqi town on different days. Dark color represents the high density of taxi location; taxi drivers are more likely to go to Yunqi on this special day. (a) October 4, 2017. (b) October 11, 2017.

A large number of advantages of human mobility in real estate investment and urban planning attract many scholars. Fu *et al.* [1] applied the ClusRanking framework to rank the value of the real estate by considering human mobility and geographic dependencies. Li *et al.* [2] designed and improved the method of neural collaborative filtering for restaurant site recommendation. Han *et al.* [3] explicitly utilized similarity with contextual information to improve the POI recommendation accuracy. Chen *et al.* [4] proposed a fast adaptively weighted matrix factorization to solve the problem of empirical bias. Chen *et al.* [5] assumed that the existing public toilets are

Manuscript received December 18, 2020; accepted January 11, 2021. Date of publication January 14, 2021; date of current version September 29, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 62073295 and Grant 62072409, in part by the Zhejiang Provincial Natural Science Foundation under Grant LR21F020003, and in part by Fundamental Research Funds for the Provincial Universities of Zhejiang under Grant RF-B2020001. Paper no. TII-20-5649. *(Corresponding author: Xiangjie Kong.)*

The authors are with the College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China (e-mail: gjshen1975@zjut.edu.cn; zhenzhenzhao_97@outlook.com; xjkong@ieee.org).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TII.2021.3051934.

Digital Object Identifier 10.1109/TII.2021.3051934

reasonable, and Chen *et al.* [6] presumed that the established bike-sharing stations are right. They all trained a classification model with supervised learning and used the model to guide the building of public facilities. These works contribute and motivate us a lot in recommending the place location, but there are still two main shortcomings. One is that most models need the same category labels of existed places for training. Due to the unreasonable buildings in early urban planning, it is quite difficult to ensure that all the labels are correct. The other is that some models can only work on historical addresses because of the cold start problem.

To the best of our knowledge, we design an unsupervised framework for commercial district discovery. First, we regard the region derived from the road segment as a node. The distribution of POIs in the road region can be the edge characteristic as the reflection of geography. The taxi trajectory location in the road region can be the node characteristic as the reflection of popularity. Second, we apply the clustering algorithm in different human mobility patterns and get intersections of different clusters. Finally, we use the results of kernel density estimation (KDE) to tell suitable commercial sites. To conclude, our article's contributions are as follows.

1) We propose an unsupervised commercial district discovery framework via embedding space clustering on graph convolution networks (GCN2CDD) to address commercial district discovery by mining different human mobility patterns.
2) We propose an embedding space clustering method to mine the candidate commercial district, which considers human mobility and geographic similarity at the same time.
3) We evaluate the proposed framework with the data of Xiaoshan District of Hangzhou City and parts of Chengdu City. The results demonstrate the effectiveness of the framework.

The rest of this article is organized as follows. Section II introduces the related work about human mobility and graph convolution networks (GCNs). Section III presents some preliminary concepts and an overview of our framework GCN2CDD. Furthermore, we introduce our methods in detail in Section IV. Section V describes our experiment settings and verifies the effectiveness of the proposed framework. Finally, Section VI concludes this article.

II. Related Work

Two main parts of the commercial district discovery are popularity and geography. Besides, commercial sites are more likely located near the road. In the popularity dimension, taxies have the characters of working all time and cover the main areas of one city. Thus, taxi trajectory data can reflect the demand of human beings, that is to say, we can use taxi trajectory data to model the popularity of a region. In the geography dimension, we can obtain the distribution of main POI categories, bus stations, and metro stations around a road region. Then, the crucial problem is how to combine popularity and geography. GCNs own the ability to extract features on non-Euclidean distance by defining the node and edge feature matrices. Then, it can aggregate the node features depending on the edge features. Thus, GCNs can fit the problem well. Next, we review the related work about human mobility and GCNs.

A. Human Mobility

Understanding human mobility is crucial for epidemic control, traffic forecasting, and, more recently, various mobile and network applications [7]. In the epidemic control field, Fang *et al.* [8] quantified the impact of lockdown on human mobility and the spread of infectious viruses. In the meantime, they proved the effectiveness of enhanced social distancing policies. In the aspect of mobile and networks, human mobility can be seen how people work in social networks. Kong *et al.* [9] designed a novel framework to optimize edge cooperative networks in wearable Internet of things and verified the framework's availability on explaining how soccer players cooperate. In the domain of traffic forecasting, Geng *et al.* [10] improved the ride-hailing demand forecasting accuracy by considering the human mobility between different city regions. These applications all acquire excellent results and demonstrate the crucial importance of human mobility. More specifically, urban human mobility indicates how people move in cities, such as driving to the workplace, going shopping, and taking a taxi to the destination. Thus, we have to consider human mobility in discovering commercial districts. Human mobility reflects the different functions and demand in a city. Yuan *et al.* [11] described the relationship between mobility patterns and POIs by feeding the POI feature vector and mobility into a Dirichlet-multinomial-regression-based topic model. This job can help us understand different functional regions in a city and human behaviors by observing directly on human flow. On the other hand, we may not get an accurate number of people's flow because of privacy protection. Thus, taxi trajectory data is a reasonable substitute to some degree. Lu *et al.* [12] succeeded in inferring POI lifetime status by applying the taxi trajectory data. Motivated by these jobs, we try to use the taxi trajectory data to model urban human mobility and discover the commercial districts in this article.

B. Graph Convolution Networks

It is well known that convolutional neural networks have great power of extracting features on Euclidean distance, for example, the images. Motivated by the idea of convolution kernel, many scholars pay attention to apply the convolution operation to graph-structured data, which result in the so-called GCNs. Spectral-based GCNs own a solid mathematical foundation in the field of signal processing [13]. They assume that the graphs are undirected and apply Fourier transform to convert the convolution operation into product operation. For an undirected graph, it is evident that the normalized graph Laplacian matrix L is symmetric and positive semidefinite. The normalized graph Laplacian matrix can be factored as $L = U\Lambda U^T$ by applying eigenvalue decomposition. Thus, the spectral graph convolution of graph signal x with a filter g_θ is defined as Hadamard product $g_\theta * x$. Nevertheless, the computational complexity is high. Therefore, many kinds of research attach importance to

simplifying complexity. One representative work is that Kipf and Welling [14] first introduced the fast approximate convolution on the graph with layerwise propagation rule for semisupervised node classification. The GCN, a typical graph neural network (GNN) model, got the favor of many areas rapidly with the advantage of abstracting the feature on the graph data. Yan *et al.* [15] adopted the GCNs in skeleton-based action recognition; they use the sampling function to construct different neighbor nodes. On the other hand, it is not suitable for the traffic flow forecasting problem because the proposed GCNs depend on the static adjacency matrix consisting of 0 or 1. Thus, Guo *et al.* [16] combined the GCNs and the attention mechanism to design the dynamic adjacency matrix to obtain the dynamic spatial information of the road network. Although they all acquired wonderful results, these applications of GCNs need a large number of given labels to ensure the performance. Thus, Sun *et al.* [17] proposed the multistage self-supervised GCNs by the idea of deep clustering to alleviate the question. Moreover, GCNs do not work well when the node's neighbors are dissimilar [18]. These works inspired us a lot in our framework. However, for the commercial district discovery problem, there are still several issues that need to be further explored, for example, how to construct the adjacency matrix and feature matrix by considering urban human mobility and geography at the same time.

III. Overview

In this section, we introduce some notations, definitions, concepts, and an overview of the proposed framework.

A. Preliminary

Definition 2.1 (GPS point): A GPS point is a position on the earth, which is usually represented as 2-D coordinates (longitude, latitude). There are many coordinate systems to describe the GPS point, such as WGS-84, BD09, GCJ02, and Mercator. We first need to transform the GPS point from different sources into the same coordinate systems. Then, we can do geometric calculations.

Definition 2.2 (POI): A POI represents a specific location indicating the property of the land use, where human beings will take activities. It is also described as a point on a map. Every POI includes information about name, coordinate, and category. The number of POI usually indicates the intensity in a district.

Definition 2.3 (Road region): One district can be divided into many small cells by major roads. A road segment is a line consisting of two points and is the basic element of the road network. To get the small cells, we add a little deviation to the road segment. Then, a road region is nearly a circle taking the offset road segment as the center. We partition the researched district into many road regions, and each road region is indexed by a unique ID.

Definition 2.4 (Taxi trajectory): A taxi trajectory is the driven route of a taxi, which is represented as a sequence of GPS points with a time slot. It also contains the basic information about the taxi, such as plate number, speed of the taxi, and the status of the taxi.

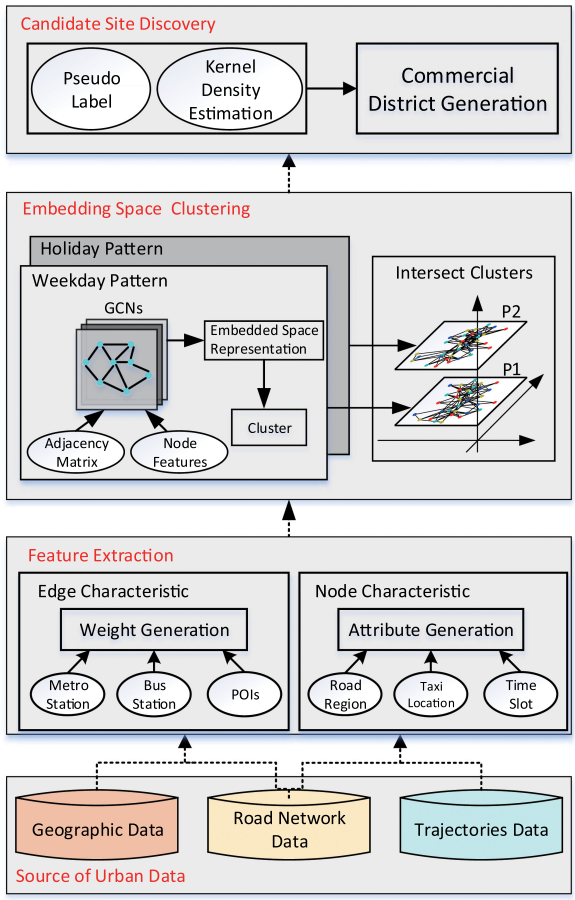


Fig. 2. Overview of the proposed framework.

Definition 2.5 (Commercial district): A commercial district is a region that owns many POIs; people will come there in leisure time or shopping time. With a direct or indirect commercial value, it plays an important part in a city's economic development.

Problem definition: Given the road network data (R), city-wide POI data (Q), and taxi trajectory (Tr) data, we aim at finding the road regions (Rr) that is similar to the commercial district.

B. Framework

We propose a framework to determine appropriate commercial sites by using heterogeneous urban data; the main structure of our framework is illustrated in Fig. 2. The three main parts are data feature extraction, embedding space clustering, and candidate site discovery. First, we generate the road regions according to part of the Xiaoshan and Chengdu District's road segments and regard the road region as a node in the graph. Then, we construct the node and edge characteristics from heterogeneous urban data. Furthermore, we use GCNs to generate a representation in embedding space and apply the clustering algorithm to get pseudo labels for different kinds of road region. Finally, we obtain the commercial district by the calculated KDE results. All these steps will be described in detail in the next section.

IV. Methodology

In this section, we elaborate on feature extraction, embedding space clustering, and candidate site discovery of the proposed framework shown in Fig. 2.

A. Feature Extraction

To determine suitable commercial districts, we need to adopt mobility and geographic similarity to GCNs in our framework. There are three main steps: data preparation, node characteristic construction, and edge characteristic construction. In the next, we will elaborate details about these steps, and the pseudocode of feature extraction is shown in Algorithm 1.

1) Data Preparation: First, we have bidirectional road segments, which can distinguish the different driving directions, in the Xiaoshan and Chengdu road networks. However, the pairwise road segments are unnecessary because we cannot build commerce in the middle of the two road segments. If the two road segments share the common intersection, we denote the two road segments to be the same. Second, we roughly choose the main urban area and discard the inconvenient transportation area. Third, we choose the major road in the remaining road network and add a little deviation to the road segment to ensure that the road segment is in the middle of the road region. Finally, we clean the dirty and wrong taxi trajectory data. The trajectory data of taxis can be divided into two categories: one is the trajectory data generated in the occupied status, and the other is the trajectory data generated in the nonoccupied status. The latter cannot directly reflect human mobility, especially on our researched problem. Thus, we only keep the data of the former. Specifically, we perform statistics analysis on the taxi trajectory data. Then, we delete the anomalous and incorrect data according to the analysis. Furthermore, we delete the taxi trajectory data that are not in the study area. In the end, we match the remaining taxi trajectory data to the road network.

2) Node Characteristic Construction: We generate the road region from the road network data in the data preparation step. Considering that people go to the nearest commercial site by walking, the road region is nearly a circle at the center of the road segment. The radius of the circle is 800 m, which is the distance of human walking in about 10 min. Then, we do statistics about taxi trajectory data to identify the number of taxies' locations in each road region per hour. The amount of taxi trajectory data can reflect the popularity of this road region; human will not take a taxi to the dissatisfied areas. Finally, we can get the node characteristic representation; it can also be called the node feature matrix X.

3) Edge Characteristic Construction: Based on the road region, we analyze the number of each POI category, bus station, and metro station. Then, we can see the geographic factors as a vector. If the two road regions own alike numbers of geographic factors, they can be though as one functional category. Motivated by the works on measuring the similarity of two distribution [19], we use the Pearson correlation coefficient to analyze the geographic influence among any two road regions. The higher score of Pearson correlation, the more similar the geographic distribution. Given the two vectors U and V, the Pearson correlation coefficient calculates the score r by the following formula:

Algorithm 1: Feature Extraction.

Input: Road Network Data R, City-wide POI Data Q and Taxi Trajectory Data Tr;
Output: Features Matrix X,Adjacency Matrix A;

initialize $X = \varnothing$, $A = \varnothing$
R, Q, Tr = **prepare_data**$\{R, Q, Tr\}$
for *each road region* R_i *in* R **do**
  for *hour* h_j *in one day* **do**
    calculate the number of taxi GPS points N_j
    $X[j, i] = N_j$
  end
  calculate one-dimensional geographic factor V_i
  $A[i, :] = V_i$
end
A = **pearson_correlation**(A)
if $A[i, j] < 0$ **then**
  Update $A[i, j] = 0$
else
  break
end
$X = StandardScaler(X)$

$$r = \frac{\sum_{i=1}^{n} (U_i - \bar{U})(V_i - \bar{V})}{\sqrt{\sum_{i=1}^{n} (U_i - \bar{U})^2} \sqrt{\sum_{i=1}^{n} (V_i - \bar{V})^2}} \tag{1}$$

where $\bar{U}$ and $\bar{V}$ are the averages of the value in vectors U and V, respectively. n is the dimension of the vector.

We analyze the same POI categories, bus stations, and metro stations to ensure that the geographic vectors have the same dimension; then, we can adopt the formula. The reason why we choose the Pearson correlation is that its value ranges from -1 to 1. The results can help us know whether the two vectors are positive correlation or negative correlation. To construct a sparse matrix, we set the negative r equal to 0. Finally, we obtain a 2-D adjacency matrix A.

B. Embedding Space Clustering

Human beings taking activities act in different patterns on different days; we divide the data into two categories according to holidays and weekdays. To obtain the features of road regions, we apply GCNs to combine the node and edge characteristics for each type of data. With the GCNs, we can map the nodes' features into embedding space, where the nodes can be divided into different clusters easily. The process of GCN embedding space is shown in Fig. 3. Given an adjacency matrix A and the feature matrix X, the GCN model constructs a filter on the graph, which will capture the node features by the edge features and get new node representations. To reduce the calculation complexity, the Chebyshev Polynomial approximation [20] graph convolution can be rewritten as

$$g_\theta * x = \sum_{k=0}^{K} \theta_k T_k \left(\frac{2L^{\text{sys}}}{\lambda_{\max} - I_N} \right) x \tag{2}$$

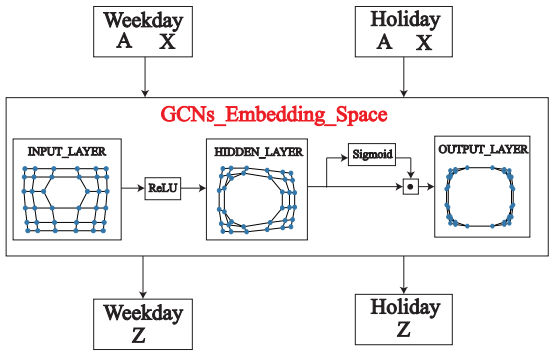


Fig. 3. Process of GCN embedding space.

where $L^{\text{sys}} = I_N - D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$, I_N is the identity matrix, D is the degree matrix, $D = \sum_j A_{ij}$, $\lambda_{\max}$ is the largest eigenvalue of L^{sys}, $T_k(\bullet)$ is the Chebyshev polynomial of order k, and θ is the vector of polynomial coefficients. Therefore, the layerwise graph convolution can be rewritten as

$$H^{l+1} = \sigma\left(g_\theta H^l W^l + b^l\right) \tag{3}$$

where H^l is the output of l layer, W^l, b^l is the parameter of l layer, and $\sigma(\bullet)$ is the sigmoid activation function for a nonlinear model. Motivated by the work in [21], we design a two-layer GCN model to embed the features in our framework, which can be expressed as

$$f(A, X) = (g_\theta H_0 W_1 + b_1 + H_0) * \sigma\left(g_\theta H_0 W_1 + b_1\right) \tag{4}$$

where $H_0 = \text{Relu}(g_\theta X W_0 + b_0 + X)$ denotes the output of the first layer with residual connection; g_θ can be calculated according to formula (2) in the preprocessing step. Then, we get the node representation in embedding space, where the nodes aggregate the geographic similarity nodes' features. To get an excellent implicit representation in GCN embedding space, we first feed the node features and adjacency matrix of graph into GNN, and then, we can get the implicit space vector expression. Furthermore, we use inner product to reconstruct the original adjacency matrix. Finally, we optimize the parameters of GCNs by comparing the difference between the reconstructed adjacency matrix and the original adjacency matrix. The loss function defined as

$$L = \text{E}_{q(Z)}\left[\log_p(A|Z) - \text{KL}[q\left(M(Z)\right)||p\left(S(Z)\right)]\right] \tag{5}$$

where $\text{KL}[q(\bullet)||p(\bullet)]$ is the Kullback–Leibler divergence between $q(\bullet)$ and $p(\bullet)$, $M(Z)$ and $S(Z)$ denotes the mean vector and standard deviation vector of Z, respectively, and Z is the output of GCNs.

With the representation in GCN embedding space, we apply the hierarchical clustering (HC) on each different type of data, and the final results are their intersections. The reason why we choose HC is that it can discover the hierarchical relationship between different clusters from the raw data. Furthermore, it does not need to define the number of clusters in advance, and its clustering rule is easily made. Now, because we want to discover the commercial districts from different road regions, the hierarchical relationship is important, and we cannot know the number of functional regions in a city with few prior knowledge. Thus, the HC can fit this characteristic well. The detailed clustering procedure is presented in Algorithm 2.

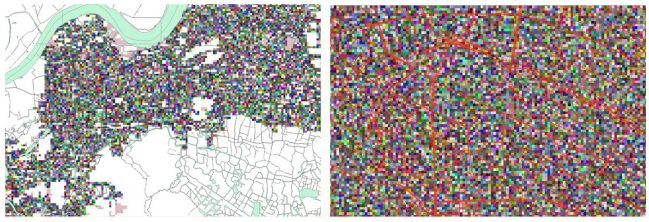


Fig. 4. Overview of fine-grained KDE results. Each color represents the KDE score of fine-grained regions. (a) Xiaoshan district. (b) Chengdu district.

Algorithm 2: Embedding Space Clustering.

Input: Features Matrix X, Adjacency Matrix A;
Output: Intersecting clusters $InterClus$;

```
initialize the GCNs parameters
initialize Z1 = ∅, Z2 = ∅
initialize clusters = ∅, InterClus = ∅
Z1 = GCNs(weekday(A1,X1))
Z2 = GCNs(holiday(A2,X2))
for Z in {Z1,Z2} do
    initialize threshold as α, sim_max as MAX
    while sim_max lager than α do
        initialize sim = ∅
        for each two nodes n_i,n_j in Z do
            sim.append(correlation(n_i,n_j))
        end
        sim_max, i, j = find_max(sim)
        n_k = aggregate(n_i,n_j))
    end
    clusters.append(n_k)
end
InterClus = intersect(clusters)
```

C. Candidate Site Discovery

We have obtained the intersecting clusters by analyzing different human patterns. Then, the crucial problem is how to give each cluster a semantic label. In other words, we need to tell whether the pseudo label is suitable for commerce. Inspired by the work in [11] and the property of commerce, we choose all the shopping POIs in the researched district and apply KDE in fine-grained regions. Given n points in 2-D spatial space, the KDE estimates the intensity by a kernel density estimator

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{6}$$

where h is the bandwidth and $K(\bullet)$ is the Gaussian kernel function

$$K(x) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}x^2\right). \tag{7}$$

Then, we can get the global fine-grained KDE results for the whole Xiaoshan and Chengdu shopping POI. The results are shown in Fig. 4. If fine-grained regions intersect with the road

TABLE I
STATISTICS OF DATASETS ABOUT EXPERIMENTS

DataSets	Properties	Xiaoshan	Chengdu
Road Networks	Number of road segments	51050	*
	Researched major roads	1312	6388
	Researched road regions	656	3194
POI	Number of POI	53015	69843
	Percentage of shopping	38.0%	19.3%
	Percentage of company	15.4%	13.6%
	Percentage of restaurant	9.8%	13.4%
	Percentage of hotel	1.5%	1.9%
	Percentage of school	1%	0.3%
	Percentage of hospital	0.2%	2.6%
	Percentage of residence	2.2%	17.8%
	Percentage of sight	0.5%	0.6%
	Percentage of entertainment	3.2%	3.9%
Taxi Trajectories	Number of taxies	8964	*
	Time period	Sept.,25 - Oct.,17	Oct.,1 - Oct.,16
	year	2017	2018
	Number of GPS points	≈ 5G per day	≈ 4G per day
	Effective days	16	16
Bus Station	Number of bus stations	22855	13832
Metro Station	Number of metro stations	218	48
Commercial Centers	Number of commercial centers	21	24

The symbol ∗ means that we cannot know the specific number.

regions, we calculate the average score of these fine-grained regions. Finally, we can get the KDE score for each road region. If road regions in each cluster have a high KDE score of shopping POI and the high popularity, then it might be the commercial districts.

V. EXPERIMENTS

In this section, we first introduce the datasets used in our article and experimental settings. Then, we conduct extensive experiments to evaluate the effectiveness of the proposed framework.

A. Hardware and Software Environments

The experiments are conducted on a computer with 64-GB memory, a Intel Xeon Silver 4110/2.1 GHz CPU, and a Quadro M4000/8G GPU. Moreover, the proposed approach and all neural-network-based baseline models are implemented based on PyTorch 1.5.0 with the cuda101 using the Python language 3.6.5.

B. Data Description

The multiple urban data contain road network data, POI data, and taxi trajectory data. More specific information about the datasets is shown in Table I.

1) Xiaoshan Datasets: The road network data cover the main area of the Xiaoshan District of Hangzhou City, China. The bus stations, metro stations, and taxi trajectory data cover the whole of Hangzhou City. The POI data are provided by Beijing GISUNI Information Technology Company, Ltd., which is a company that aims at geographic science. The commercial centers are crawled from Baidu Maps, which is the largest online map service platform in China.

2) Chengdu Datasets: The road network data cover the part area of Chengdu City, China. The taxi trajectory data are collected by the GAIA Open Dataset relying on DiDi company, which is the most popular mobile transportation platform in China. The POI data, bus stations, metro stations, and commercial centers are crawled from Baidu Maps, which is the largest online map service platform in China.

TABLE II
RESULTS ABOUT SUPERVISED ALGORITHMS AND OUR ALGORITHMS ON XIAOSHAN DATASETS

Algorithms	Measurement		
	Precision	F1-score	Recall
AdaBoost	0.556	0.503	0.524
DT	0.389	0.477	0.589
LR	0.662	**0.582**	0.596
RF	**0.665**	0.549	0.489
SVM	0.660	0.386	0.408
GCN2CDD	**0.431**	**0.520**	**0.655**

C. Experimental Settings

First, taxi trajectory data are divided into two categories according to holidays and weekdays. Then, we calculate the average value of the number of taxis' locations per hour in the same category and take the value as the final results. Because of the urban data from different data sensors, we need to transfer the coordinates into Mercator projection plane coordinates to count the number of each POI category, metro, and bus station of the road region. To evaluate the effectiveness of our methods, we define the commercial center with a radius of 800 m as a commercial district. If the road region intersects with commercial districts, we mark the road region as a commercial area. The inputs of the proposed framework are the researched road regions, and the outputs of the proposed framework are the road regions marked as commercial districts. Thus, we use the following three metrics, i.e., recall, precision, and F1-score, to evaluate the proposed framework:

1) *recall:* the proportion of correctly classified samples in all samples;
2) *precision:* the proportion of true positive samples in positive samples predicted by the model;
3) F1-score: comprehensively measured the values of precision and recall.

To verify the validity of the unsupervised framework, we investigate several common supervised learning algorithms to observe these algorithms' performance on our data, including logistic regression (LR), decision tree (DT), random forest (RF), adaptive boost (AdaBoost), and support vector machine (SVM). For all these machining learning models, we optimize the parameters with the cross validation. The experimental results of supervised learning algorithms and our algorithm on different datasets are summarized in Tables II and III. According to the experimental results, our framework owns the highest recall score, and the other two metrics are relatively low on Xiaoshan datasets. One possible reason is that the road regions in selected Xiaoshan Datasets are sparse, which may result in notable features. Even if the features are gathered by GCNs, it is also easy to distinguish. Thus, the effect of our framework on this dataset is not particularly obvious. However, on the Chengdu datasets,

TABLE III
RESULTS ABOUT SUPERVISED ALGORITHMS AND OUR ALGORITHMS ON CHENGDU DATASETS

Algorithms	Measurement		
	Precision	F1-score	Recall
AdaBoost	0.470	0.536	**0.623**
DT	0.404	0.460	0.535
LR	0.382	0.338	0.305
RF	0.457	0.494	0.538
SVM	**0.504**	0.523	0.544
GCN2CDD	**0.503**	**0.554**	**0.616**

our framework owns the highest F1-score, and the other two indicators are almost close to those of the supervised method. Therefore, our unsupervised model can achieve similar effects to the basic supervised method. In the case of without labels, our algorithm will have an obvious advantage.

Furthermore, we also design some contrasted methods in our proposed framework to reveal the crucial part of GCNs.

1) *Without_GCNs:* We directly apply the HC algorithm on holiday and weekday node feature matrices and get the intersection of clusters.
2) *Multilayer perceptron (MLP) [22]:* This model works well in learning latent features and is approved effective in neural collaborative filtering, which is a classic recommend algorithm. We use the MLP model with two hidden layers and choose the rectified linear unit activation function.
3) *LSTM/GRU [23]:* These two models perform excellently in learning the sequential features. We treat the road regions as a sequence and adopt the model to fit our framework.

To express the influence of the GCN layer, we conducted comparative experiments on different GCN layers. Thus, we carried out experiments separately on single-layer (**GCNs_1**), two-layer GCNs (**GCNs_2**), and three-layer GCNs (**GCNs_3**). We initialize all the parameters in these models with Kaiming uniform distribution, and the negative slope of the rectifier is set to $\sqrt{5}$. Furthermore, these models' parameters are optimized by the idea of an autoencoder [24]; then, we can get the latent representation. The threshold of the HC algorithm is set to the optimal value for all the model; we will discuss in detail how we select the threshold in the next section. We treat the road regions in the cluster that has a high KDE score of shopping POI and the high popularity as the commercial districts. Then, we can finish the precision, F1-score, and recall measurements, and the results are shown in Fig. 5.

The GCNs_2 model outperforms other algorithms in terms of precision and F1-score measurement. However, the GCNs_3 model performs worse than the GCNs_2 model. Since the embedding of nodes is already very similar, the deep layers do not always perform well [18], [25]. The framework without any embedding methods owns a higher score of recall but a lower score of other measurements. It means that it identifies most of the road regions as commercial districts, which is very unreasonable. The MLP embedding methods cannot capture the

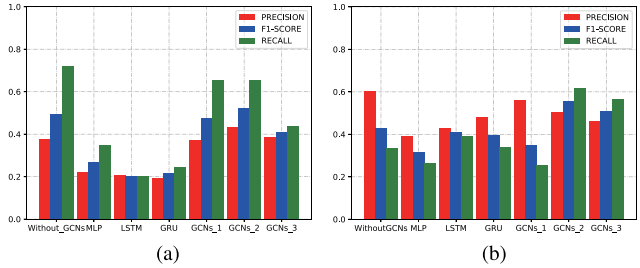


Fig. 5. Results of measurements on different methods. (a) Results on Xiaoshan datasets. (b) Results on Chengdu datasets.

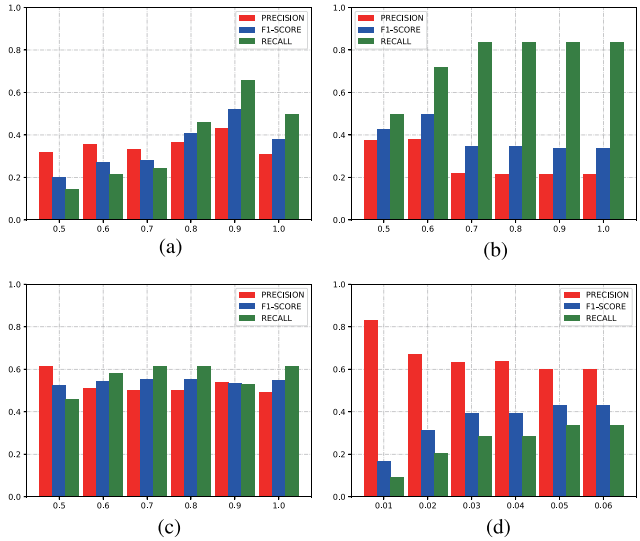


Fig. 6. Measurements of different threshold values. (a) GCNs_2 on Xiaoshan datasets. (b) Without_GCNs on Xiaoshan datasets. (c) GCNs_2 on Chengdu datasets. (d) Without_GCNs on Chengdu datasets.

logical similarity of different road regions. Therefore, it cannot work well in our framework. The LSTM/GRU embedding methods perform badly because the road regions lack serial correlation. The experiments are evidence of the fact that GCNs can outperform the other neural networks when extracting features with an unsupervised mode. The improvement of our proposed framework is mainly attributed to adopting human mobility and geographic similarity into GCNs. In this way, our framework can discover the functional regions hidden in the different human mobility patterns.

To get the best threshold of the HC algorithm, we set the threshold range as 0.5–1 with the step size of 0.1. Given examples of Without_GCNs and GCNs_2 models, the experimental results on Xiaoshan and Chengdu datasets are shown in Fig. 6. Experiments show that GCNs_2 and Without_GCNs models perform best when the threshold is 0.9 and 0.6 on Xiaoshan datasets, respectively. Moreover, GCNs_2 and Without_GCNs models perform best when the threshold is 0.8 and 0.05 on Chengdu datasets, respectively. On the one hand, the Without_GCNs model owns a high precision score but low recall score, which means that this model cannot recognize all the positive samples. On the other hand, without the aggregation characteristics of GCNs, the threshold is hard to find. Although

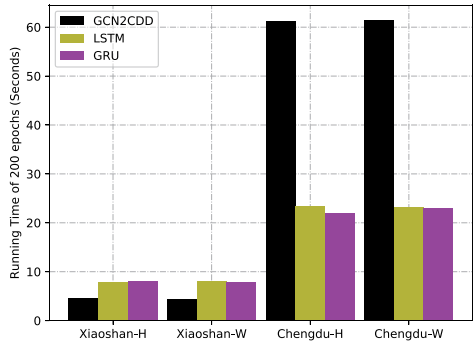


Fig. 7. Running time comparison. -H and -W denote the holiday pattern data and weekday pattern data, respectively.

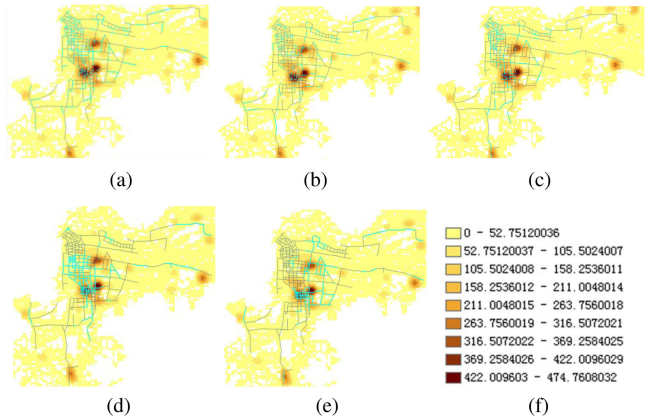


Fig. 8. HHrs results of different neural networks embedding space. (a) MLP results. (b) LSTM results. (c) GRU results. (d) Without_GCNs. (e) GCNs_2. (f) represents the legend of map.

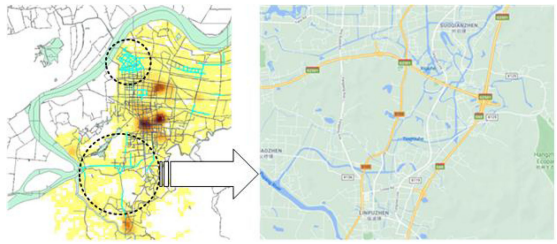


Fig. 9. Detail information of HLrs. The yellow line in the right map represents the arteries.

we try to enlarge the threshold of more than 0.06 to improve the performance, we can only get one clustering result. This explains that the model cannot distinguish any category.

Considering the actual deployment issues, we analyzed the time complexity of the framework. The experimental results are shown in Fig. 7. In the Xiaoshan datasets, the pretraining time required by our framework is relatively short. However, in the Chengdu datasets with more road regions, the pretraining time increases sharply. Thus, the experiments show that the time complexity of our method is affected by the number of road regions to be detected. However, we can apply the framework in the mode of cloud edge collaboration. Specifically, we pretrain the model parameters in the cloud platform. Then, we can distribute the parameters to the edge servers to complete the deployment of the entire framework.

D. Discussion

To better understand the semantic label for each cluster, we visualize the Xiaoshan KDE results of shopping POI and analyze the rationality. Thus, we will analyze in detail some representative experimental results in this section. The road regions with less popularity may need further development, and we do not consider these road regions temporarily. In this article, we define the road regions into three main categories: High popularity with the Low density of shopping POI road regions (**HLrs**), High popularity with the Middle density of shopping POI road regions (**HMrs**), and High popularity with the High density of shopping POI road regions (**HHrs**). Then, we use the ability to distinguish these three categories as the evaluation criteria for model performance. We display the HHrs results (marked as blue in the pictures) of the framework under different conditions in Fig. 8.

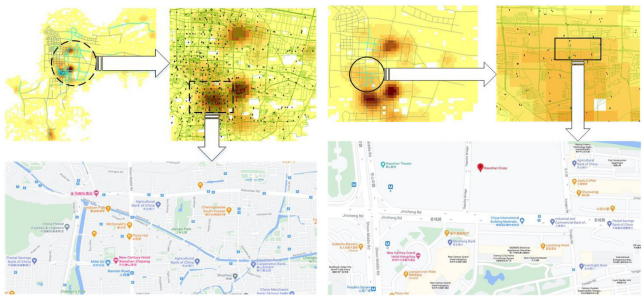
The results in Fig. 8(a)–(c) show that the areas marked as HHrs are scattered in different parts of the study area and cannot cover the dark areas in the pictures. It means that they cannot distinguish the HHrs well. Thus, we do not discuss these models in detail.

Then, we select the best performance of Without_GCNs and GCNs_2 and choose the HHrs results, which are shown in Fig. 8(d) and (e), to explain the crucial importance of GCNs. The high-density shopping POI road regions (represent as dark color) cannot be group together, and the light-colored road regions and dark-colored road regions are more likely classified as the same category in the Without_GCNs model, which represents that the Without_GCNs model cannot tell HHrs and HMrs well. In contrast, the results of the GCNs_2 model can cover most dark-colored areas. It means that the GCNs_2 model can distinguish the HHrs well even if the HHrs are not close in the physical road network. To sum up, without the information aggregation in GCN embedding space, the boundary between HMrs and HHrs is hard to find.

We display the detail of the three categories of GCNs_2 and try to give the semantic labels. Furthermore, the results detected by our framework are all marked as blue in pictures. As shown in Fig. 9, HLrs are mainly the connection between two shopping centers or other Districts in Hangzhou City. HLrs can be thought of as the important road regions because human beings have to go through these regions to arrive at the destination.

As displayed in Fig. 10(a), the left picture explains the overview of HHrs results, and the right picture shows the detail of selected regions. HHrs own many shopping malls, famous Chinese restaurants, and convenient transportation, as shown at the bottom of Fig. 10(a); people are more likely to take these regions as the destination in shopping time. Thus, HHrs can be thought of as developed commercial districts.

HMrs also own convenient transportation, a certain number of shopping malls, and other services. Meanwhile, the color in the map shows that HMrs have lower POI density than HHrs.



(a) (b)

Fig. 10. Detail information on Google Maps. The black spots represent the bus stations, and the green five-pointed stars denote the metro stations. (a) Detail information of HHrs. (b) Detail information of HMrs.

Thus, HMrs are more suitable for commerce development and can be thought of as developing commercial districts. The successful case is the new open Xiaoshan Cross, which is a famous large-scale shopping mall, in 2018, as shown in Fig. 10(b). People in Xiaoshan are more likely to go to Xiaoshan Cross shopping centers. To conclude, our framework can distinguish the important road regions, developed commercial districts, and developing commercial districts well. Thus, it can guide merchants to choose commercial sites with less labor and time.

VI. Conclusion

In this article, we proposed a data-driven framework for commercial district discovery based on different urban data sources, such as taxi trajectory data, road network data, and POI data. The framework learns the features from the GCN embedding space, which considers the geographic and human mobility's influence. Then, the HC algorithm was applied to mine different human mobility patterns to get pseudo labels. According to the pseudo labels and KDE, we obtained the potential commercial districts. We finally evaluated the effectiveness of the proposed framework on the Xiaoshan District of Hangzhou City and parts of Chengdu City.

References

[1] Y. Fu, H. Xiong, Y. Ge, Z. Yao, Y. Zheng, and Z.-H. Zhou, "Exploiting geographic dependencies for real estate appraisal: A mutual perspective of ranking and clustering," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2014, pp. 1047–1056.

[2] N. Li, B. Guo, Y. Liu, Y. Jing, Y. Ouyang, and Z. Yu, "Commercial site recommendation based on neural collaborative filtering," in *Proc. ACM Int. Joint Conf. Int. Symp. Pervasive Ubiquitous Comput. Wearable Comput.*, 2018, pp. 138–141.

[3] P. Han *et al.*, "Contextualized point-of-interest recommendation," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, 2020, pp. 2484–2490.

[4] J. Chen *et al.*, "Fast adaptively weighted matrix factorization for recommendation with implicit feedback," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 3470–3477.

[5] C. Chen, Y. Liu, C. Liao, C. Chen, L. Feng, and Z. Wang, "Where to build new public toilets? Multi-source urban data tell the truth," in *Proc. IEEE SmartWorld, Ubiquitous Intell. Comput., Adv. Trusted Comput., Scalable Comput. Commun., Cloud Big Data Comput., Internet People Smart City Innov.*, 2019, pp. 1162–1169.

[6] L. Chen *et al.*, "Bike sharing station placement leveraging heterogeneous urban open data," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, 2015, pp. 571–575.

[7] K. Zhao, S. Tarkoma, S. Liu, and H. Vo, "Urban human mobility data mining: An overview," in *Proc. IEEE Int. Conf. Big Data.*, 2016, pp. 1911–1920.

[8] H. Fang, L. Wang, and Y. Yang, "Human mobility restrictions and the spread of the Novel Coronavirus (2019-nCoV) in China," *Nat. Bur. Econ. Res.*, Cambridge, MA, USA, Working Paper 26906, 2020.

[9] X. Kong *et al.*, "Mobile edge cooperation optimization for wearable Internet of things: A network representation-based framework," *IEEE Trans. Ind. Informat.*, to be published, doi: 10.1109/TII.2020.3016037.

[10] X. Geng *et al.*, "Spatiotemporal multi-graph convolution network for ride-hailing demand forecasting," in *Proc. Conf. Artif. Intell.*, 2019, pp. 3656–3663.

[11] J. Yuan, Y. Zheng, and X. Xie, "Discovering regions of different functions in a city using human mobility and POIs," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2012, pp. 186–194.

[12] X. Lu, Z. Yu, C. Liu, Y. Liu, H. Xiong, and B. Guo, "Inferring lifetime status of point-of-interest: A multitask multiclass approach," *ACM Trans. Knowl. Discov. Data*, vol. 14, no. 1, pp. 1–27, 2020.

[13] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *IEEE Signal Process. Mag.*, vol. 30, no. 3, pp. 83–98, May 2013.

[14] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. Int. Conf. Learn. Representations*, 2017.

[15] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, 2018.

[16] S. Guo, Y. Lin, N. Feng, C. Song, and H. Wan, "Attention based spatial-temporal graph convolutional networks for traffic flow forecasting," in *Proc. Conf. Artif. Intell.*, 2019, pp. 922–929.

[17] K. Sun, Z. Lin, and Z. Zhu, "Multi-stage self-supervised learning for graph convolutional networks on graphs with few labeled nodes," in *Proc. Conf. Artif. Intell.*, 2020, pp. 5892–5899.

[18] Y. Xie, S. Li, C. Yang, R. C.-W. Wong, and J.Han, "When do GNNs work: Understanding and improving neighborhood aggregation," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, 2020, pp. 1303–1309.

[19] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Item-based collaborative filtering recommendation algorithms," in *Proc. 10th Int. Conf. World Wide Web*, 2001, pp. 285–295.

[20] D. K. Hammond, P. Vandergheynst, and R. Gribonval, "Wavelets on graphs via spectral graph theory," *Appl. Comput. Harmon. Anal.*, vol. 30, no. 2, pp. 129–150, 2011.

[21] B. Yu, H. Yin, and Z. Zhu, "Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, 2018, pp. 3634–3640.

[22] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua, "Neural collaborative filtering," in *Proc. 26th Int. Conf. World Wide Web*, 2017, pp. 173–182.

[23] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in *Proc. NIPS 2014 Workshop Deep Learn.*, 2014.

[24] T. N. Kipf and M. Welling, "Variational graph auto-encoders," in *Proc. NIPS Workshop Bayesian Deep Learn.*, 2016.

[25] G. Li, M. Muller, A. Thabet, and B. Ghanem, "DeepGCNs: Can GCNs go as deep as CNNs," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 9267–9276.